sidequest

Drafted by Adelin Cai
Last edited: April 6, 2021

Building a Trust & Safety Mindset

This document is intended to be a high level guide for folks looking to foster a trust and safety function within their organizations. Targeted at early stage organization leadership and product builders who don't have prior experience in trust and safety work, this guide answers some frequently asked questions.

What is trust and safety?

The work of preventing people from being harmed while using online services has existed ever since the first groups of people started interacting with one another online. From reviewing and taking action on abusive speech and graphic violence, to setting up systems preventing spam, phishing, and fraud, people have been tasked with ensuring that online services are free from harmful content and behavior that could jeopardize the health of the user community.

eBay is credited with having first used the term "trust and safety" when setting up the team that ensured the integrity of the service. Today, many teams use the term "trust and safety" to describe the entire function of safeguarding the online user experience. While some organizations opt to use other team names like "Platform Integrity" or "Anti-Evil Operations", the fundamental goals of these teams are very much the same: They work to "develop and enforce principles and policies that define acceptable behavior and content online[1]" at the user-facing level and behind the scenes as a result of the online service's product experience.

You can think of trust and safety work as generally preventing these (sometimes overlapping) categories of content and behavior:

Illegal or Regulated	Harmful	Disruptive
Refers to content and behavior that are restricted as a result of local laws and regulations. Examples include: • Intellectual property infringement • Distribution of Child Sexual Exploitation Material (CSAM) • Promotion of illicit drugs • Processing payments from countries under sanction • Access to products or services based on age	Refers to content and behavior that are directly damaging to people's sense of safety, and the health of the community. Examples include: • Violent threats and doxxing • Misinformation • Discrimination on the basis of race, gender, sexual orientation, etc • Fraudulent payments and other transactions resulting in financial loss	Refers to content and behavior that undermine the quality or value of the service. Examples include: • Spamming with unwanted content • Manipulating systems to artificially change service statistics and data • Any other content or behavior that doesn't align with the mission of the service

[1] See the Trust & Safety Professional Association (TSPA). Disclosure: I am a co-founder of the organization, and serve as the non-compensated chair of its board.

Drafted by Adelin Cai
Last edited: April 6, 2021

The non-exhaustive breakdown below illustrates some areas of expertise/functions within the trust and safety industry. These categories aren't mutually exclusive:

Content review operations	Full time content reviewers
	Contractor/fixed term content reviewers
	Specialized content reviewers (violent extremism, terrorism, hate speech, child sexual exploitation, influence operations, etc)
	Advertiser content reviewers
	Marketplace/merchant content reviewers
	Advertiser and merchant account reviewers
Legal requests and law enforcement response	Intellectual property specialists
	Government takedown requests specialists
	User data requests specialists
	Law enforcement requests specialists
Content policy development and creation	Non-monetized content policy specialists (violent extremism, terrorism, hate speech, etc)
	Monetized content policy specialists (regulated goods and services)
Spam, fraud, account compromise, and payment risk operations	Spam, phishing, account takeover, malware specialists
	Payment fraud and risk specialists

Why should I think about trust and safety?

If your online service is intended for use by a third party (users who generate content or otherwise interact with an element of your service, like search results, advertisers who pay for distribution, entities that outsource services to be fulfilled by you, etc), you have a responsibility to protect people from harm as a result of using your service, and protect your online service from activity that could compromise its long term health and viability.

For every positive interaction that you build for a third party, there's a risk of simultaneously facilitating unacceptable content or behavior. Proactively considering and mitigating harm and abuse are fundamental to your long term growth, and is an integral part of your product's and team's success.

In practical terms, a proactive mindset about your organization's trust and safety efficiency could translate into cost savings for reputation management, crisis comms, and even lobbying or litigation. In terms of organizational health, maintaining the integrity of your product or service could contribute directly into increasing user growth and user retention.

Want to read up on different trust and safety cases? Here's a library of real-life trust and safety dilemmas that illustrate the challenges that you may have to face, depending on the scope of your product or business: Trust & Safety Foundation Case Studies[2]

When do I need to start thinking about trust and safety?

Right now. No, really. You should start incorporating trust and safety commitments as soon as you start thinking about what your online service provides, and begin to develop your product or service. Think of trust and safety as a function or mindset that is central to the core user experience, not a separate post-facto solution for when things go awry.

If you're already past the product/service development phase, it's still not too late to be mainstreaming trust and safety principles in your organization's work. Go through the exercise of (re-)allocating your organization's finite resources to ensure that you're taking into account the need to build a responsible and long term trust and safety infrastructure. The more time and resources you spend considering and solving risk mitigation upfront, the lesser the impact of the inevitable abuse, and the more responsive the clean up when it occurs.

How can I set up this team for success?

To start, promote a team culture of authenticity and vulnerability, and internalize the importance of work-life balance. This is crucial in developing an organization where members feel psychologically safe enough to communicate transparently about their needs.

The most impactful process to support trust and safety teams is simple. On a regular basis:

1. Understand and familiarize yourself with the work the team does
2. Proactively and directly ask the team which resources they need, and what they think you could be doing to make their work better and easier
3. Actively listen to the responses from the team, and internalize the feedback
4. To the best of your ability, fulfill the team's needs

Other things you could do to elevate the team's work as a core part of your organization's infrastructure:

- Believe in and explicitly communicate the importance of trust and safety work as a core part of the user experience, and its contributions to your organization's growth
- Commit impactful technical and operational resources to help the team achieve measurable success

[2] Disclosure: I am a co-founder of the organization, and serve as a non-compensated member of its board.

- This work can be extremely difficult, and at its worst can result in a range of adverse effects from emotional exhaustion, depression, physical ailments, suicide ideation, and more. Commit to the team's health and safety by providing accessible, professionally administered, and evidence-based wellness and resilience support
- Encourage and support the trust and safety team's engagement with the broader trust and safety community (For example, by keeping current on the Trust & Safety Professional Association's resources, events, and workshops)

With these pointers for beginners, you're well on your way to building a trust and safety mindset within your organization! Don't delay, the time is now.

Acknowledgements
This work would not be possible without the guidance and expertise of the following trust and safety leaders and colleagues I've been fortunate to work with: Jen Clarke, Karl Conway, Michelle Dy, Jaymi Green, Del Harvey, Jud Hoffman, Alana Karen, Micaela McDonald, Julie Mora-Blanco, Janett Riebe, Charlotte Willner, Morgan Woodward, and so many more who continue to do stellar work.